Introducing HDInsight

In this video, we are going to cover several HDInsight concepts, and how Azure provides added benefits that you can take advantage of.
Azure HDInsight deploys and
provisions managed Apache Hadoop clusters in the cloud,
providing a software framework designed to process, analyze and
report on big data with high reliability and availability.
HDInsight uses the Hortonworks Data Platform or
HDP Hadoop distribution.
Hadoop often refers to the entire Hadoop
ecosystem of components.
Which includes Apache HBase, Apache Spark, and Apache Storm
as well as other technologies under the Hadoop umbrellas.
You can deploy these technologies on Windows or
Linux clusters and can spin up the cluster in minutes.
You can process semi-structured and unstructured data and
develop in Java, .NET, Python, and more.
You can also visualize your data in Excel, Tableau or Power BI.
And it integrates with the following,
Azure Blob Storage and Azure Data Lake Store for
Hadoop file system, otherwise known as HDFS.
Azure Portal for management and administration.
And then you can also use Visual Studio for
application development tooling.
With HDInsight, deploy Hadoop in the cloud without buying new
hardware or incurring other upfront costs.
There's also no time consuming installation or
setup, Azure does it for you.
Launch your first cluster in minutes.
Azure HDInsight provides a software framework,
designed to process, analyze and
report on big data with highly reliability and availability.
Meaning data reuse and sharing, elastic scale-out,
lower data storage costs, and protection against data loss.
The high speed flat networks in Azure data centers provide fast
access between the virtual machines in the cluster and
Azure Blob Storage or Azure Data Lake Store.
So data movement is very efficient.
HDInsight also integrates with BI tools such as Power BI,
Excel, SQL Server analysis services and
SQL Server reporting services.
Here is the Hadoop ecosystem.
Hbase is a non-relational database written in Java.
It uses HDFS as its file system.
This database is used in scenarios where we need to store

sparse data, small information found within large data.
For example,
finding the 50 largest items in 2 billion objects.
Facebook's messaging platform uses an HBase database.
Pig is is a high level platform for
creating MapReduce jobs for Hadoop.
It also uses a language called Pig Latin,
which is very similar to SQL.
It can be extended further to be used from other programming languages like JavaScript.
Pig was developed by Yahoo and
then moved to the Apache Software foundation in 2006.
Hive is developed by Facebook and
it helps in providing BI capabilities on top of Hadoop.
We can analyze Hadoop's HDFS data using Hive's querying language which is called HiveQL,
which is also similar to SQL in syntax.
HCatalog is a storage management layer for Hadoop.
Basically, it helps in presenting HDFS data in
relational format.
Polybase allows SQL Server PDW users to execute queries against data stored in Hadoop's HDFS.
So as an example, users can file queries that join data between HDFS and PDW tables.
Mahout is a set of machine learning algorithms
that can use Hadoop for processing.
Oozie is a server based workflow engine specialized in running workflow jobs with actions that run Hadoop Mapreduce and
Pig jobs.
Flume is used to collect, aggregate and
move large amounts of log data.
Scoop is used to transfer data between relational and Hadoop.
Microsoft uses a scoop based connector
to move data between SQL Server and Hadoop.
The Hive ODBC driver is used to connect
different Microsoft products to Hive.
Which in turn provide connectivity with Hadoop.
Products like Excel, PowerPivot,
SharePoint Insights, SSRS, SSaS etc.
REST products can connect to Hadoop's HDFS using REST APIs as well.
A common requirement for this scenario is from small devices or from apps running outside of a Hadoop cluster,
where Hadoop's native programming is not available.
This REST connectivity is possible through implementation
of web HDFS.

Microsoft Insight clusters can be managed using the system center management pack for HDInsight.
HDInsight is a 100% Apache Hadoop based service in the cloud,
and a Hadoop distribution powered by the cloud, this means HDInsight was architected to handle any amount of data.
Scaling from terabytes to petabytes on demand.
You can spin up any number of nodes any time.
We charge only for
the compute and storage that you actually use.
You can get immersive insight using HDInsight with the familiar BI tools, such as Excel, PowerPivot SQL Server analysis services and SQL Server reporting services.
To retrieve, analyze and report data integrated with HDInsight by using either the Power Query add-in or
the Microsoft hive ODBC driver.
How is Microsoft contributing to Hadoop?
We're starting at the bottom in the source code,
to make those source base work better, not just for Windows, but for everyone.
Some of Microsoft's community contributions are,
tens of thousands of lines of code contributed across all deliverables.
6,000 plus engineering hours contributed since February 2012.
Other contributions are Apache build verification and infrastructure, working with the Apache infrastructure team and Hadoop core PMC on donating Azure VMs to be used as Jenkins servers for continuous integration.
Interactive query, contributing code and
query processing experience to help with Hive query performance such as Stinger, ORC and Tez projects.
Hadoop on Windows in versions 1.0 and 2.0 contributed back to our porting efforts for Hadoop on Windows including command line scripts for the Hadoop surface area.
Mapping HDFS permissions model to Windows,
Native Task controller for
Windows, implementation of Hadoop native libraries for Windows such as compression codecs and Native I/O.
ASV driver contributed our file system implementation for Azure storage.
Microsoft is super engaged with contributors and committers.
We've logged 6,000 engineering hours committed to Stinger project work, doing work on security integration and ensuring Hadoop works great on Windows.
Microsoft and
Hortonworks promote open Hadoop in various ways.

For instance, we have joint engineering teams where
we share people on each other's teams.
They report to Hortonworks and Hortonworks report to us.
Our joint engineering team also shares resources.
This is not just a re-licensing of a distro,
we are one engineering team.
Also, we've worked to increase the milestone cadence.
We do a 12 month calendar year roadmap with each other.
That roadmap is what each company locks to,
revisiting it every three months.
When Cloudera released Impala, Microsoft and Hortonworks
changed the resources to start working on the Stinger project.
In calendar year 13, this became a joint commitment to build Tez,
which is a vectorized query executor, and Stinger.
Joint planning and execution joint sign-offs, Microsoft and
Hortonworks signed off on Linux, Windows, and Azure.
Aligned offerings, HDP on Windows HDI appliance and
HDI service.
Plus there's a joint roadmap and calendar year planning,
roadmap developments, plus quarterly adjustments.
Hortonworks is making Hadoop an enterprise viable platform,
look at the partner page on the Hortonworks site.
Engineering alignments, Microsoft works with Hortonworks
to develop Hortonworks data platform for Windows.
HDInsight on Windows Azure and
HDInsight on PDW is based on HDP.
Microsoft and Hortonworks, are working on various Hadoop
sub-projects together like Stinger and Tez.
And have contributed over 6000 engineering hours and
25,000 plus lines of code.
Then there is corporate alignment.
Microsoft is one of Hortonworks strategic partners,
joint marketing alignment,
such as webinars, event strategy, press, etc.
Plus joint support for alignment.
Microsoft provides level one support and pays Hortonworks for
level two and level three.
There's also field alignment.
Hortonworks field reps will get quota compensation or
relief if HDP for Windows, or if PDW on Azure is sold.
Microsoft field reps work with Hortonworks reps,
to get credibility from the Hadoop story.
There are reasons why HDInsight is the right customer choice for
deployment on either Windows or Linux.
As you may know, you can select Linux or Windows clusters when
deploying big data workloads into Microsoft Azure.

With Windows, leveraging existing Windows based code including .NET to scale over all of your data in Azure.
With Linux, customers can more easily move existing Hadoop workloads into the cloud and incorporate additional big data components to run in the service.
By offering a choice for Windows and Linux clusters, Microsoft is enhancing flexibility for customers to create insight from the massive amounts of data being created in the cloud with the OS of their choice.
Apache Hive is data warehouse software, built on Hadoop, that allows you to query and manage large datasets and distributed storage using a SQL-like language called HiveQL.
Hive is an abstraction on top of MapReduce.
When run, Hive translates queries into a series of MapReduce jobs.
You can also improve the performance dramatically with Stinger and Apache Tez combined with Hive.
Microsoft is starting to contribute to the Stinger initiative to dramatically speed up Hive and make it more enterprise ready.
Our main activity in Stinger right now is around Tez and vectorized query execution.
Tez is a framework that allows data intensive applications such as Hive, to run much more efficiently at scale.
In the latest release of HDInsight, Hive supports running on Tez.
Tez is enabled by default for Linux based HDInsight clusters.
Apache HBase is an open-source NoSQL database that is built on Hadoop and modelled after Google Bigtable.
HBase provides random access to strong consistency for large amounts of unstructured and semi-structured data in a schema list database organized by column families.
HDInsight HBase is offered as a managed cluster that is integrated into the Azure environment.
The clusters are configured to store data directly in Azure Blob Storage, which provides low latency and increased elasticity in performance and cost choices.
Apache Mahout is a scalable library of machine learning algorithms that run on Hadoop.
Using principles of statistics, machine learning applications teach systems to learn from data, and use past outcomes to determine future behavior.
Mahout contains algorithms for processing data such as filtering, classification and clustering.
HDInsight support also includes Apache Storm which is an open source stream analytics platform

that can process real time events at large scale.
This lets you process millions of events as they're being generated, enabling use cases like Internet of Things, or IoT, and gaining insights from your connected devices or web-triggered events.
Apache Storm on HDInsight allows you to create distributed, real-time stream analytic solutions in the Azure environment by using Apache Hadoop.
The Azure Cloud makes Apache Storm easy and cost effective to deploy with no hardware to buy, no software to configure, your choice of development tools, Java or C# and deep integration into Visual Studio.
It consumes millions of real-time events from a scalable event broker like Apache Kafka or Azure Event Hub.
This bring near real-time analytics capabilities to HDInsight.
You can use the language of your choice.
It provides support for Storm components written in Java, C#, and Python.
It supports a mix of programming languages.
So you can read data using Java, then process it using C#, if you want.
This diagram describes an end-to-end applicability of HDInsight Storm.
Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big data analysis applications, SQL, streaming data, and machine learning.
Spark delivers queries up to 100 times faster than traditional big data queries.
It provides a common execution model for tasks like ETL, batch queries, interactive queries, real time streaming, machine learning, and graph processing on data stored in Azure storage.
It makes interactive queries over big data in Hadoop using BI tools or open source notebooks.
You can create a streaming solution for IoT or a real time application.
Use machine learning algorithms to be able to predict outcomes in your analysis with the notebook experience of Jupyter or Zeppelin.
You can customize Hadoop clusters to run other Hadoop projects.
This Hadoop ecosystem is a portfolio of fast moving open source projects that are evolving quickly.
To give customers flexibility, HDInsight has the option to deploy arbitrary Hadoop projects through custom scripts.

This includes popular projects, like Spark, R, Giraffe,
and Solr.
Learn how to use HDInsight tools for Visual Studio to connect to
HDInsight clusters and submit Hive queries.
It is time-consuming to test Hive script against
an HDInsight cluster, taking several minutes or more.
HDInsight tools for Visual Studio is capable of validating
Hive script locally without connecting to a live cluster.
HDInsight tools for Visual Studio also enables users to see
what's inside the Hive job by collecting and
surfacing the yarn logs of certain Hive jobs.
You can also use Visual Studio to visualize Hadoop clusters,
tables, and storage.
Hive editor supports IntelliSense.
HDInsight tools for Visual Studio, supports loading
the remote metadata when you are editing your Hive script.
For example, when you type select * from,
the IntelliSense lists all the suggested table names.
When a table name is specified,
the column names are listed by IntelliSense.
The tool supports almost all Hive DML statements,
sub queries and the built in UDFs.
Azure HDInsight is an Apache Hadoop distribution,
powered by the cloud.
This means that it handles any amount of data,
scaling from terabytes to petabytes on demand.
Spin up any number of nodes at any time.
We charge only for the commute and storage you use.
Microsoft notes that more than half of
the Fortune 500 are among the quarter of a million customers
now using Azure.
Microsoft is boosting capability at an even faster rate
than customer growth.
Doubling its Azure compute and storage capacity every 6 to 9
months, along with expansions in Japan, Australia, and
its 21Vianet operated Chinese Azure Cloud.
In the past year, Microsoft has doubled its Azure customer base
to 250,000, and averages around a 1,000 new customers per day.
With over 4 trillion objects in Windows Azure, and
an average of 270,000 requests per second,
customer requirements grow and are met daily.
In peak periods,
demand can grow to a staggering 880,000 requests per second.
This diagram shows that Azure easily integrates with your
existing IT environment through the largest network of secure
private connections, hybrid databases, and storage

solutions, and data residency and encryption features.
So your assets stay right where you need them.
And, with Azure Stack, you can bring the Azure model of
application development and deployment to your data center.
Azure hybrid cloud solutions give you the best of
both worlds.
More IT options, less complexity and cost, this is why it is one
of the best cloud computing services available.
With HDInsight, you can deploy Hadoop in the cloud without
buying new hardware or other upfront costs.
There is also no time consuming installation or setup,
Azure does it for you.
You can launch your first cluster in minutes.
HDInsight in the cloud bypasses capacity planning.
It handles any amount of data,
scaling from terabytes to petabytes on demand.
Spin up any number of nodes at any time.
We charge only for the compute and storage that you use.
Azure is mission critical and enterprise ready.
With the automatic provisioning of data clusters,
HDInsight clusters are much easier to create than
manually configuring Hadoop clusters.
This is Managed Hadoop, backed by an SLA,
you can get three nines of availability and 99.9% uptime.
High availability and reliability of clusters,
a second headode has been added to the Hadoop clusters, deployed
by HDInsight to increase the availability of this service.
Standard implementations of Hadoop clusters typically have
a single headnode.
HDInsight removes the single point of failure
with the addition of a secondary headnode.
The switch to a new high availability cluster
configuration doesn't change the price of the cluster,
unless customers create clusters from an extra large headnote
instead of the default large size headnote.
Azure handles all maintenance tasks in HDInsight so
that you can focus on your strategic work.
OS patching and
security updates are performed automatically by Azure requiring
minimal IT resources for upgrades or patching.
Hadoop versions are rapidly releasing throughout the year.
So you are always on the latest version of Hadoop
without effort.
This helps minimize IT resources to update Hadoop versions.
Open source does not mean free.
Deploying Hadoop on-premises still requires servers and

skilled Hadoop experts to set up, tune and maintain them.
A cloud service lets you spin up a Hadoop cluster in minutes
without upfront costs.
In the Microsoft Azure Cloud, HDInsight is billed by usage.
You only pay for
what compute and storage that you use when you use it.
Spin up a Hadoop cluster, analyze your data, and
then shut it down to stop the meter.
Clusters can be deleted when no longer used.
Create a Hadoop cluster in minutes, and
add notes on demand.
The cloud offers organizations immediate time to value.
Azure HDInsight combines the best of Hadoop
open source technology with the security, elasticity, and
manageability that enterprises require.
We have built it to integrate with Excel and Power BI,
our business intelligence offering, allowing people
to easily connect to data through HDInsight,
and then refine and do business analytics in a turn key fashion.
Excel is the BI tool for everyone.
Already, 1 billion Microsoft Office users can connect to
HDInsight to analyze and visualize data.
Power BI helps with collaboration and
new experiences for Share, Ask and Access data.
Azure machine learning not only provides tools to
power predictive analytics, but also provides a fully managed
service you can use to deploy your predictive models,
as ready to consume web services.
Azure Machine Learning provides tools for creating complete
predictive analytics solutions in the cloud.
Quickly create, test, operationalize, and
manage predictive models.
You do not need to buy any hardware, nor
manually manage virtual machines.
Azure Machine Learning has everything you need to create
predictive analytics solutions in the cloud from a large
algorithm library, to a studio for building models,
to an easy way to deploy your model as a web service.